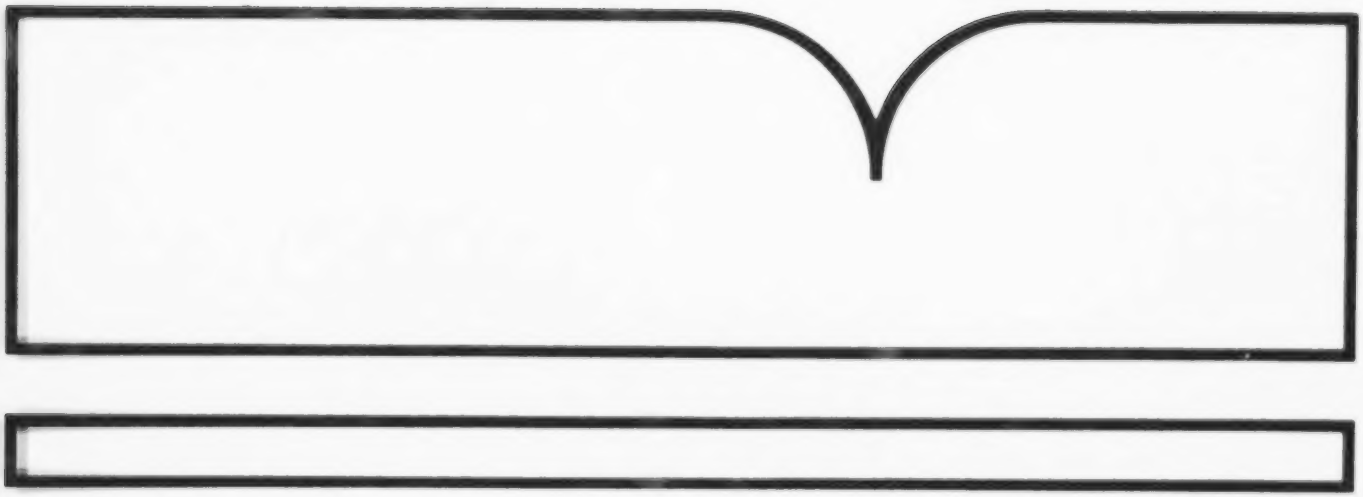
PB87-170478

Application of Cluster Analysis to Aerometric Data (Journal Version)

(U.S.) Environmental Monitoring Systems Lab.
Research Triangle Park, NC

1986



U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before complet

1. REPORT NO. EPA/600/J-86/251	2.
3.	PB87-170478
4. TITLE AND SUBTITLE APPLICATION OF CLUSTER ANALYSIS TO AEROMETRIC DATA (Journal Version)	5. REPORT DATE 1986 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) H.L. Crutcher, R.C. Rhodes, M.E. Graves, B. Fairbairn, and A.C. Nelson	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS H.L. Crutcher, Consultant, 35 Westall Ave., Asheville, NC 28804 R.C. Rhodes, USEPA, MD-77B, RTP, NC 27711 M.E. Graves, Northrop Services, RTP, NC 27709 B. Fairbairn, PEI Associates, 505 S Duke St., Durham, NC 27701 A.C. Nelson, UNC @ Wilmington, Wilmington, NC 28406	10. PROGRAM ELEMENT NO. 11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS U.S. Environmental Protection Agency Office of Research and Development Environmental Monitoring Systems Lab/Quality Assurance Div. Research Triangle Park, NC 27711	13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA/600-08

15. SUPPLEMENTARY NOTES
J. of the Air Pollution Cont[illegible] Association, October 1986, Vol. 36, No. 10, pp. 1116-1122.

16. ABSTRACT

The NORMIX data analysis progra[illegible]ch incorporates cluster analysis and multivariate statistical analysis routines, was modified and revised for use in a UNIVAC 1110 computer. The revised program was tested on three sample data sets and produced results in agreement with those from the original program. The NORMIX program was then used to evaluate and analyze eight sets of aerometric data from various sources. Comparison of the performance of NORMIX with two other cluster analysis algorithms, MIKCA and SAS CLUSTER, revealed that all three programs produce similar results in terms of hierarchical clustering, but NORMIX produces considerably more statistical evaluation and information to the user. Thus NORMIX is recommended as the most useful cluster analysis program of these three.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field Group

REPRODUCED BY
U.S. DEPARTMENT OF COMMERCE
NATIONAL TECHNICAL
INFORMATION SERVICE
SPRINGFIELD, VA 22161

18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS (This Report) UNCLASSIFIED 20. SECURITY CLASS (This page) UNCLASSIFIED	21. NO. OF PAGES 8 22. PRICE

EPA/600/J-86/251
JOURNAL ARTICLE

TECHNICAL PAPERS

JAPCA 36: 1116–1122 (1986)

Application of Cluster Analysis to Aerometric Data

Harold L. Crutcher[1], Raymond C. Rhodes[2], Maurice E. Graves[3], Beth Fairbairn[4], and A. Carl Nelson[5]

[1] Consultant, Asheville, N.C.
[2] U.S. Environmental Protection Agency, Research Triangle Park, N.C.
[3] Northrop Services, Inc., Research Triangle Park, N.C.
[4] PEI Associates, Inc., Durham, N.C.
[5] University of North Carolina, Wilmington, N.C.

The NORMIX data analysis program, which incorporates cluster analysis and multivariate statistical analysis routines, was modified and revised for use in a UNIVAC 1110 computer. The revised program was tested on three sample data sets and produced results in agreement with those from the original program. The NORMIX program was then used to evaluate and analyze eight sets of aerometric data from various sources. Comparison of the performance of NORMIX with two other cluster analysis algorithms, MIKCA and SAS CLUSTER, revealed that all three programs produce similar results in terms of hierarchical clustering, but NORMIX produces considerably more statistical evaluation and information to the user. Thus NORMIX is recommended as the most useful cluster analysis program of these three.

Environmental data are being collected for research and monitoring purposes at an ever-increasing rate. Many techniques are used to analyze the data. In many cases, the data are adequately analyzed or evaluated using simple and traditional methods considering the behavior of single variates or the relationships between a small number of variates. However, with the collection of data for a larger number of concomitant variates and with the increased capability of electronic computers, numerous techniques of multivariate analysis are being applied to evaluate data and to determine patterns and relationships of the data in *n*-dimensional space. Multivariate techniques, including principal component, factor, discriminant function, and regression analysis, were evolved in other disciplines but are being modified and extended for use in environmental disciplines.

Clusters of univariate and bivariate data can easily be detected visually from graphical plottings and clusters of trivariate data can with more difficulty be detected by 3-dimensional plots. Recently-developed computer graphics can enhance the detection of trivariate clusters. Clusters in more than three dimensions, however, are almost impossible to visualize and can be detected abstractly only through the use of rather extensive and complicated cluster analysis. In recent years, cluster analysis and other types of multivariate analysis have been effectively used to detect predominant combinations of chemical pollutant species and/or associated meteorological data that are identified with specific point sources or other types of sources. The application just mentioned, however, is but one example of how cluster analysis may be effectively used in analyzing environmental data.

It is most effective to use some types of cluster analysis in conjunction with other multivariate methods. Furthermore, in many cases it is logical initially to perform cluster analysis and then perform other multivariate analyses on each of the identified clusters. One reason is that outlier values or small outlying clusters are easily identified by most cluster analysis techniques. All multivariate methods are sensitive to outlier data, whether these data are valid or invalid. Thus, some of these techniques can be used to edit and quality assure data to produce relatively clean data banks.

A prerequisite for discriminant function analysis, linear or nonlinear, is that some prior identification of clusters or subgroups of data observations must be made. If identification of the clusters or subgroups is not previously made on the basis of prior scientific information, the cluster analysis identifies clusters from the data themselves.

In principal component, factor, or regression analyses, if several clusters of data lie along a given *n*-dimensional axis, a routine application of these techniques may not reveal the different clusters which may well have particular scientific meaning. Each of the clusters of data may have a basically different internal structure or pattern which can be explored by individual principal component, factor or regression analysis on each cluster. Such differences in internal structure of the individual clusters will not be easily detected by application of principal component, factor or regression analysis alone in the first data pass.

Various investigations in many fields have developed in totality a multitude of algorithms to define and detect clusters of data. It was considered desirable in this study to compare several of the different available clustering algorithms in their application to environmental data. The models or algorithms used were NORMIX, MIKCA, and SAS CLUSTER.

Objective

Production of valid aerometric data bases requires adequate instrumentation and maintenance, representative exposure, competent personnel, excellent communications, and a fully sufficient quality assurance and control system in an ongoing observational program. To aid in the development of valid data bases and to extend the methods of data analysis, five specific goals of a study on aerometric data clustering were defined:

1. Convert and adapt Wolfe's NOR-

Copyright 1986–Air Pollution Control Association

NTIS is authorized to reproduce and sell this report. Permission for further reproduction must be obtained from the copyright owner.

MIX clustering program for use on EPA's 1110 UNIVAC computer.
2. Expand the program to preview the data, incorporate transformations with back transformations, and add various graphical displays.
3. Calibrate and validate the converted and expanded program.
4. Demonstrate the application of the computer program to various types of environmental data.

These goals have been met and documented, as discussed later.

General Comments

Data banks of aerometric and associated information range from small to large and from limited to extensive in their coverage. Randerson[1] edits the Department of Energy publication, "Atmospheric Sciences and Power Production." This is new and contains much background material for those working in the Atmospheric Sciences discipline. Hunt *et al.*[2] discuss two applications of quality assurance techniques to aerometric data. With the increasing interactive problems of man and the environment, as more information is required on the state of the environment, data collection installations are beginning to produce prodigious amounts of information. In order to make sense of these data banks or to understand the salient aspects of the information becoming available, considerable effort must be expended to utilize the available exploratory techniques or to develop new ones. Principal components,[3-9] factor,[7-15] cluster,[15-26] discriminant function or pattern recognition,[16,17,23,27] and regression[28-30] analysis techniques exist.

All the techniques attempt, in one way or another, to condense the extensive information into small packets or units which carry the requisite core information, yet physically and conceptually take up less space. Data transformations aid in all of these techniques. Though not necessarily true or necessary in all of the techniques, most of the techniques do consider an observed value of a variate to exist along a directional line or axis in space. With ten variables, there will be ten axes. An observational vector then consists of ten component vectors, one component provided by each variate. Principal component analysis[3-9] seeks to establish ten new axes upon which the projections of the observational vector will be a new transformed variate along each axis. Eigenvectors describe the orientation of the new axes and the eigenvalues denote the variances of the new variates. The combination of a new axis and the largest variance identifies the first principal axis. The next axis associated with the next largest variance is the second principal axis. The axis associated with the least variance will be the tenth principal axis. Thus, principal component analysis attempts to produce new variates in which the major portion of the variance resides in as few of the variates as possible.

Suitable rotation of the axes produces orthogonal axes, which implies that the correlations among the new variates are zero. This is the usual concept in use in principal component analysis. However, the user may have other concepts in mind and may not want completely orthogonal relationships determined among the new data. The rotations may be sought where the developed techniques may be used to deduce a maximum of the sought-for information. That is, new variates (axes) are sought among which maximum correlation resides. Thus, oblique axes and planes and weights may be sought and obtained. These weighting factors thus lead to the naming of this general technique using weighting factors as factor analysis.[7-15] Both the principal component and factor analysis techniques can be used to develop clustered data, though the first seeks zero correlation and the second seeks maximum correlation among the new variates. However, cluster analysis can be done without the use of principal component or factor analysis, but these can help in cluster development and interpretation of results. In particular, component or factor analysis may be most effectively applied to the individual groups first identified by cluster analysis.

Hotelling[4] makes a very important point. Though there may be ten original variates, and though principal component or factor analysis indicates that two or three of the new component variates account for the major portion of the variance with or without zero correlation, exclusion of any of the variates, with subsequent recalculation, may produce quite different results with attendant different interpretations.

How many clusters are possible in a data set? It seems advisable to provide some information on this point. Anderberg (p. 3)[19] refers to Abramowitz and Stegun (p. 824)[31] regarding the combination and permutation considerations involved as Stirling numbers of the first and second kind. For example, in the example given by Anderberg[19] in the small sample of 25 data being separated into five groups, more than two million billion possibilities exist. With the same sample but with the number of groups being allowed to vary, more than four billion billion possibilities exist. In the latter case, there might be only one group of the 25, 25 groups of one each, or any other combination or permutation thereof. The main reason that this information is proffered is to point out that the user must be content not to explore all possibilities, but to use techniques that estimate the probable cluster configurations and their characteristics in the most expedient and least costly way. Hope always exists that the space which the data presumably portray has been spanned effectively. Whether this happens will never be known except for a very small finite set for a one-time only situation which will never occur again. For all other situations, several runs of the data with different cluster seeds to initiate the clustering process may provide a general consistent concept of the cluster topography. If the data set being explored is large, it may be advisable to make a few runs on small, randomly selected data to give a broad idea as to where the cluster centroids may be. Then actual data vectors in the suspected centroid regions, or even approximate data vectors constructed by the user, may be used as cluster seeds in a final run. If the user has some previous applicable knowledge or perceptions and if this is used to construct the cluster seeds, then it is more likely that the output will be believable.

Study

Agglomerative, but not divisive, clustering was considered for this study. Wolfe,[32,33] McRae[34] and Barr[35] provide the three programs used in this study which are, respectively, NORMIX, MIKCA and SAS CLUSTER. These are initially hierarchical agglomerative programs. MIKCA iteratively adjusts the initial clusters. NORMIX has the additional capability to develop and describe the clusters as multivariate normal, with estimates of their statistical parameters, and the further capability to provide discrimination for classification purposes.

Modification and Expansion of the NORMIX Program

The NORMIX algorithm set has a substantial statistical basis with corresponding statistical assumptions and significance tests. It separates heterogeneous distributions of data into relatively homogeneous groups. For that reason, it was selected for further development in this study. The program, originally written in IBM FORTRAN IV language, was converted to the ASCII FORTRAN language requirement of the UNIVAC 1110 at NCC-EPA/RTP (Rhodes[36]). It was enlarged considerably by adding additional algorithms before and after the NORMIX algorithm set to produce needed addi-

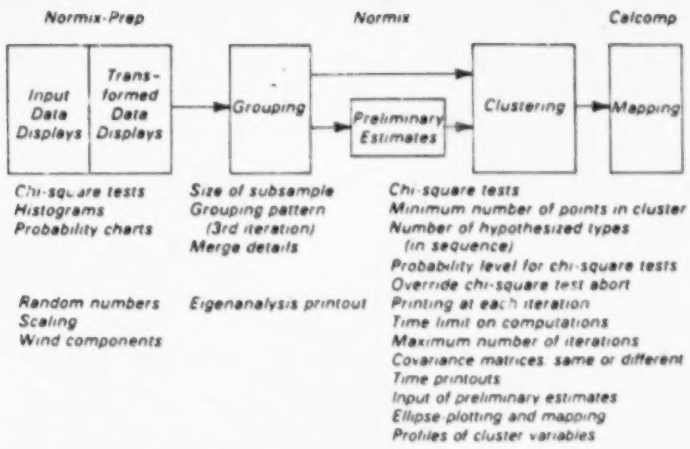


Figure 1. NORMIX flow and options.

tional exploratory information and output.

Figure 1 contains the three main sections: 1) NORMIX—PREP, 2) NORMIX, and 3) CALCOMP. The second section is the Wolfe[33] set of algorithms program modified for the UNIVAC 1110. The first and third section comprise the expansions developed in this study. The first section provides various options to examine and transform, if necessary, original input data. The third section is designed to take the NORMIX output to produce additional tables, back transformations of data, and graphical output. CALCOMP routines were used for the graphical output, hence the name. Histograms, arithmetic probability graphs, bivariate graphs and 3-sigma charts are available through use of the third section routines. The last two are illustrated later.

The first section, NORMIX—PREP, permits the user to 1) establish identification numbers for each datum or vector so that it may be traced through all operations, 2) optionally select tabular graphical output to study the distributions, variate by variate, 3) scale data and transform data, and 4) perform certain statistical tests. The acronym "NORMIX" from "NORmal MIXtures" (Wolfe[33]), implies that the output credibility depends upon the input data, appropriate transformations thereof having been sampled from clusters, each of which is approximately multinormally distributed. Crutcher[37] transforms wind as a vector with direction and speed, which are in themselves not normally distributed, to orthogonal components which are normally distributed. Thus, winds are bivariate normal in their components. Atmospheric pollutants are zero bound and may be considered to be unlimited above, enough so that the lognormal transformation may be used. (See Aitchison and Brown[38] on the lognormal distribution.) NORMIX—PREP provides for such transformations. All data are then standardized with the use of the Mahalanobis[39] distance as a metric both at this stage and in subsequent cluster treatment. Back transformations and algorithms are available in the last sections.

NORMIX—PREP allows one to examine the data to help visualize and test whether the data set, variate by variate, is unimodal and normally distributed. If all variates are unimodal and normally distributed, there are likely to be no clusters or outliers.

The second section, NORMIX, contains K-means lead-in clustering algorithms from Ward[40] and MacQueen[41]. The first is hierarchical, while the second is hierarchical only in the first stage. In the first, when a datum is once assigned to a cluster, it does not have a chance to shift to another cluster. In the second, by means of statistical estimates for each cluster, criteria are established which allow a datum to be shifted from one group to another. Using the output of the first algorithm, a first iteration produces a second set of clusters. The number of iterations is a controllable option. For this study, a second iteration is made and the result printed, yet maintained and transferred or made available as input estimates into the NORMIX algorithm proper. With the clusters so obtained, the estimated means, standard deviations, correlation matrices and mixture proportions are computed and used for the first iteration in the NORMIX algorithm. The user's control at this stage is in some previously specified options: 1) probability level to test the hypotheses; 2) difference, which is the decision criterion for convergence to final cluster determination; 3) assumption whether to use (i) equal or (ii) unequal covariances; 4) choice to have the eigenvector-eigenvalue data printed, along with the statistical characteristics of clusters at each or only the last iteration; 5) printing of subsequent discriminant function analysis classifications with classifications printed on a background of two functions at a time.

The selection and use of a specified probability level for testing really controls the number of clusters obtained. If the number of clusters specified is too small, the program halts at the specified number even though convergence has not been attained. However, the results may be quite reasonable and usable. If the number of clusters specified is too large, convergence may be attained with fewer clusters. A default option of one hundred iterations exists to cease calculations and print results to alleviate time costs.

Whenever statistical estimates of the parameters of the multivariate clusters are developed and used, overlapping of the hyperellipsoids or other hyperforms will occur. The overlapping may range from very little to much. As an example, it may be quite difficult for one to state that an individual exhibits more characteristics of the father or mother or grandparents or other relatives. One might state that the individual looks more like the maternal parentage (0.70) than the paternal parentage (0.30) or perhaps the probabilities appear to be equal (0.50, 0.50); i.e., a statement of no decision. These different assignment probabilities are perhaps not to the liking of those who want a categorical statement, for they are comfortable only with categorical statements. Perhaps to allay the fears, trepidations, doubts and insecurity of these people and to allow more people to develop a sense of reality in the multiple measurements, a new term, "fuzzy statistics," has been suggested. This may be a misnomer, but more and more literature appears each year under this name or guise. See generally the publication "CIS," Current Index to Statistics (Gentle[42]) and, specifically, Bezdek *et al.*[43] and Smets.[44] NORMIX, with input normality assured, produces viable probability statements. With five clusters, the probability assignments to multivariate normal clusters 1, 2, 3, 4, and 5 of a particular datum might be 0.40, 0.30, 0.15, 0.10 and 0.05, respectively. Another datum might have the respective assignment probabilities of 0.95, 0.02, 0.01, 0.01 and 0.01.

Two basically different approaches can be used in multivariate analysis involving pattern recognition techniques, of which cluster analysis is one method. If the approach is "learned" or "taught," the investigator uses prior knowledge or experience in guiding or directing the analysis. For example, he may know or suspect a certain number of clusters, or he may have prior estimates of the location of the centroids of the clusters, or he may know or suspect the relationships or correlation of the

variates within the clusters. In such cases, these estimates can be used as starting points and should simplify or hasten the solution. On the other hand, if the investigator wishes the data to "speak for itself," the approach is considered "unlearned" or "untaught," in which the investigator learns as the solution progresses.

Calibration and Validation of the Expanded NORMIX Program

The ability of the present modified and expanded version of the NORMIX program to produce results equivalent to those of Wolfe[33] for the same data with the older program version (and with a different computer) indicated that the new revised version had been adequately calibrated. Program validation consisted of application over several types of data sets, not necessarily all aerometric. Data validation included the isolation of outliers, if any, for examination and treatment. For treatment of outliers or suspect data, see Grubbs[46] and Barnett and Lewis.[47] Both single and clustered outliers are identifiable in the hierarchical clustering, but only clusters of at least three data are identifiable in the NORMIX processing. As the NORMIX program uses the same hierarchical clustering algorithms as several of the other available programs, it is as useful for this purpose as the others, including the MIKCA and SAS routines.

In order to validate the expanded NORMIX program, three sets of data were used:

1. A classical data set of measurements of Iris petal and sepal lengths and widths (four variates) made by Anderson[45] and used by Fisher[16] to illustrate clustering and classification. These same data (a classic set) are used to calibrate or validate most new algorithms today.
2. A synthetic bivariate data set from Wolfe,[32] and
3. A set of synthetic data consisting of three predetermined three-element data sets that could be expanded both in variances and distances among the centroids.

Sets 1 and 2 agreed with those of Wolfe. The MIKCA and SAS CLUSTER models also used the data of Set 1 for validation. The output of Set 3 returned the original stipulated clusters prior to their being mixed.

Application of the Cluster Analysis Programs

The Los Angeles Catalyst Study (LACS) data set and the extensive environmental data sets from the Community Health Air Monitoring Program (CHAMP) form the basis for this study. Rhodes[36] describes the EPA project report. Volume I presents a detailed examination of the historical development of the NORMIX algorithm (Wolfe[33]) and its application to the CHAMP data set, plus descriptions of cluster analyses of data from the Los Angeles Catalyst Study (LACS) by the SAS CLUSTER (Barr[35]) and MIKCA (McRae[34]); Symons modifies the last program. Volume II contains the modified NORMIX program with complete documentation, and Volume III contains more detailed examples and discussion of the application of the NORMIX algorithm to the CHAMP data.

Cluster analysis has been used on air pollution data to reveal cyclic patterns (over days, weeks, or seasons) and to identify local source and receptor effects. All of these relations are reflected in different combinations of values of the variables that cluster together. Because of the clustering in aerometric data, multivariate cluster analyses are useful as preliminary analytical and data validation techniques. Rhodes[36] supports the well-known inverse concentration relationship between ozone and oxides of nitrogen, presumably due to their production timing and to their chemical interaction. These and many other relationships are indicated in tabular form and illustrated in some examples by diagrams in which data are plotted with overlying distribution ellipses in bivariate form, or by profile models of means and three-sigma limits for given measurements. Simple comparison tables furnish additional general information.

Data from monitoring activities cannot be considered as random samples from a single universe, but rather the data result from sampling mixtures of distributions, usually internally correlated within each group. As expected, these studies show that pollutant data depend on, or are highly correlated with, meteorological conditions. At a given site and with a given set of pollutant sources, the pollutant concentrations at the site are heavily dependent upon the meteorological regime. Thus, the pollutant distributions will be a mixture of the distributions that result from the mixture of the meteorological regimes and the interaction of the pollutants over the time of the monitoring. Solar radiation is an effective agent in the overall regime, and in the photochemical reactions of pollutants, but these data were not available for inclusion in this study except as indirectly measured by ambient and dew point temperatures. This discussion suggests that the pollutant data distributions first be separated into meteorological regimes by cluster analysis and that these subsets then be evaluated individually by other appropriate techniques. Moreover, analysis involving prediction of future pollutant concentration distributions for each meteorological regime should also consider the probabilities of occurrence of each of the meteorological regimes. This will enhance the clustering and the classification of data.

Los Angeles Catalyst Study. In order to study the effect of the use of automobile catalytic converters on aerometric parameters, the period of the LACS data necessarily had to include the periods before and after the 1975 introduction of these devices; the period selected was 1974 through 1978. The pollutant elements (variates) observed were suspended particulates (SP), ozone (O_3), nitrogen oxide (NO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and lead (Pb). Traffic count and the meteorological variates, temperature, wind speed and wind direction, were included. For this analysis, measurements of the variates, SP, CO, Pb, wind speed, wind direction and traffic count, from two selected sites were used. These sites were on either side of the San Diego Freeway between the intersection of the freeway with Sunset and Wilshire Boulevards. At these sites, the data sets were not always complete. Effectively, in the multivariate sense, the omission of a single element requires the rejection of the entire observation from the data set under consideration. In some cases it may be reasonable to merge one incomplete multivariate vector (observation) with another complementarily incomplete vector from a nearby locale to obtain a usable complete vector. If this is done, this factor must be considered when interpreting the results.

Problems of missing or merged data were encountered in the LACS study. Using only two sites from the LACS data bank, the number of available and useful hourly observations for the 1977 through 1978 period is about 3031, as compared to about 25,000 observations originally available from all eight sites of the LACS. The reader may consult Part 2 of Volume I of the EPA Project Report (Rhodes[36]) for further details.

Community Health Air Monitoring Program

The periods of records at the CHAMP stations were relatively short, i.e., from September 1975 through November 1976 for Angwin, California and Loma Linda, California, and from August 1974 through September 1976 for Magna, Utah. The data selected for use were for certain hours, days, weeks, and seasons.

The pollutant and meteorological data consisted of oxides of nitrogen

Table I. Intercomparison of selected pollutant data throughout the year for NO, NO_x, and O_3 at three locations.

Locations	Transformed variates[a]						Number of observations
	NO		NO_x		O_3		
	Mean	S.D.	Mean	S.D.	Mean	S.D.	
Angwin, CA	0.6981	0.0017	0.7002	0.0045	0.7127	0.0087	288
Loma Linda, CA	0.7036	0.0130	−3.0442	0.9576	0.7081	0.0198	122
Magna, UT	0.7055	0.0164	0.7123	0.0216	0.7081	0.0090	324

[a] Logarithmic transformations used.

(NO_x), sample nitrogen oxide (SNO), calculated nitrogen oxide (NO), nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), total sulfur (TS), total hydrocarbons (THC), non-methane hydrocarbons (NMHC), temperature (T), dew point (DP), winds (S,W), and atmospheric pressure (P). For this study all winds were transformed to south-north and west-east components, positive from the south and west. An option in the program permits transformation of polar coordinates (wind direction and speed) to rectangular (orthogonal) coordinates, along and at right angles to any preselected direction such as might be indicated by a highway. The default option is the south-north and west-east configuration.

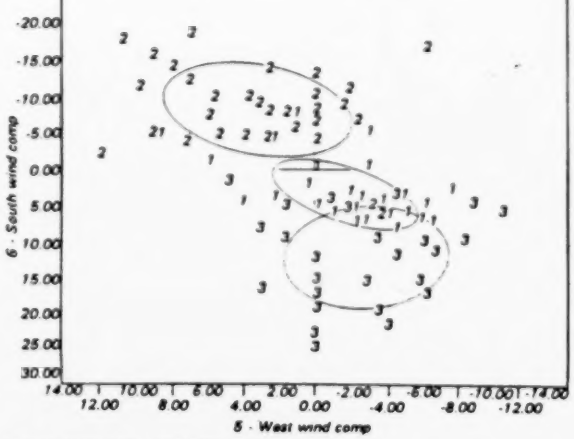
Table I shows a comparison of selected pollutant data throughout the year for NO, NO_x and O_3 at Angwin, Loma Linda, and Magna. The data were transformed logarithmically.

The information in Table I reveals that Angwin probably has the lowest levels of oxides of nitrogen and highest level of ozone, and the lowest variability of these three variates at the three stations. This is, of course, from prior informational scans, one reason why the Angwin site was selected for monitoring and for use in this study. The large standard deviation and negative mean value for the Loma Linda NO_x data reflect that the number added to low observed values (to ensure a minimum value of two before logarithms were taken) was not large enough.

Figure 2, developed in CALCOMP from NORMIX output information from Magna data, illustrates the 0.50 probability ellipses for wind speed and direction and the tabular output of the associated pollutant means, standard deviation, proportions, and the number of observations. Cluster numbers for each plotted point are included to help assess the clustering efficacy. It must be remembered that the pollutant variables have been logarithmically transformed; all decisions are made in the transformed fields. The 0.50 bivariate probability ellipses are centered on the centroids of the plots of wind components, but the cluster classifications are in terms of all variates. As previously noted, it is in the overlapping regions that errors of misclassification may occur for an individual datum. However, the statistical estimates are generally expected to provide the best estimates of the cluster configurations. This type of presentation is a projection of the total multidimensional ellipsoid onto the plane of the two selected variates. Any two variates can be selected by options provided in the program.

The program option that developed the profile plot, Figure 3, arranges the variate output in terms of the largest cluster with variate means increasing in sequence. For this case, this is Cluster 1 (triangles) with the upward rising line. In all presentations such as this, the values of the other variates follow the sequence established by the largest cluster. The upper and lower horizontal lines are the upper and lower 3-sigma limits for the processed data. Only the pollutant data were transformed. The upper and lower 3-sigma variate values are located above and below the 3-sigma lines. The physically larger numbers below the minimum three-sigma limit, ranging from 1 through 8, identify the variates in order of their entry with the program, i.e., into each observational vector. This is arbitrary and is simply the order chosen. The spread of the symbols identifying the two limits of the variates of the individual clusters permits a ready comparison of those variates which have the least or greatest variance in the cluster. Again, where transformed variates are used, the sigma limits are in the trans-

Variable	Cluster 1	2	3
1 - NO	P = .354	P = .333	P = .313
Mean	0.72	0.70	0.70
Standard deviation	0.03	0.00	0.00
2 - NOX			
Mean	0.74	0.70	0.70
Standard deviation	0.03	0.00	0.01
3 - Ozone			
Mean	0.70	0.71	0.71
Standard deviation	0.01	0.00	0.00
4 - TS			
Mean	0.78	0.72	0.71
Standard deviation	0.07	0.03	0.01
5 - West wind comp			
Mean	-1.70	3.13	-2.86
Standard deviation	3.41	4.42	3.86
6 - South wind comp			
Mean	3.40	-8.32	12.12
Standard deviation	4.12	5.56	6.00
7 - Temperature			
Mean	0.47	14.85	18.01
Standard deviation	3.49	5.59	6.22
8 - Dew point			
Mean	-5.60	-1.31	-3.83
Standard deviation	3.36	2.95	3.02



West vs South wind - probability level = 0.50

Figure 2. Magna, Utah, Weekday 3, 0.50 probability ellipses of the west-east and south-north wind components for three cluster types.

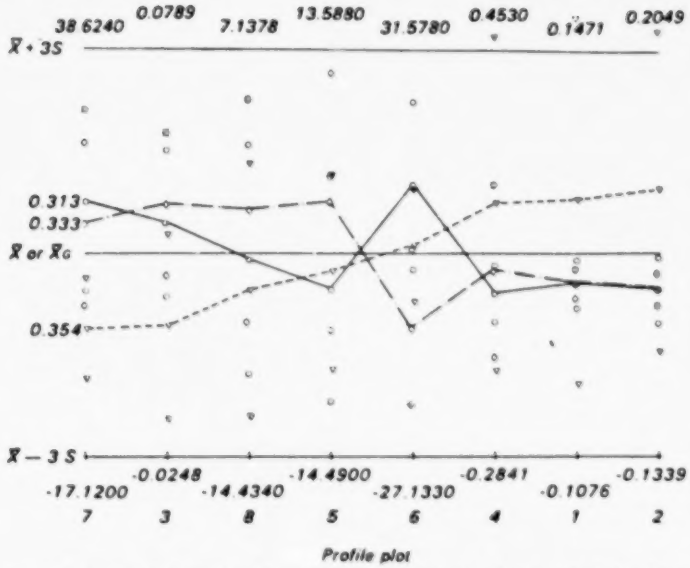


Figure 3. Magna, Utah, Weekday 3 profile plots of pollutants and meteorological variates against a three-sigma background. The means and two-sigma limits (symbols) for each variate of the three data clusters are represented by triangles (Cluster 1), diamonds (Cluster 2) and circles (Cluster 3), respectively. The variate numbers along the abscissa refer respectively to: 1, NO; 2, NO_x; 3, O_3; 4, TS; 5, *W* component of wind; 6, *S* component of wind; 7, temperature; and 8, dew point. The sequence of variates is determined by the numerical value of their means, geometric for pollutants, arithmetic otherwise, for Cluster 1 (lowest to highest). The other numbers refer to the three-sigma limit values for each variate.

formed fields. If back transformations are utilized, the representations will be asymmetric.

In Figures 2 and 3, it may be noted that weak southeast winds with a mean speed of approximately 4 km/h are associated with ozone readings lower than average and oxides of nitrogen and sulfur readings higher than average (triangles). Strong winds from the northwest, approximately 9 km/h, and from the south-southeast, approximately 12 km/h, again show the inverse relationship of ozone with oxides of nitrogen and sulfur. Clusters 1 and 3, with southeast and south-southeast winds, respectively, show the greatest temperature differences between the means, namely 17.54°C. Further investigation might yield the reason for this temperature difference. Speculatively, this feature may be associated with seasonal weather characteristics or synoptic episodes. Additionally, these features may be interpreted as an indication of source point origins as well, particularly those associated with the stronger wind, the observational point acting as a receptor point. For more details, reference may be made to Rhodes.[36] Observational posts established at or upwind to the estimated source points would help to identify source and receptor points.

Comparison of Three Clustering Programs

Table II compares three clustering algorithms, SAS CLUSTER, MIKCA and NORMIX. All three algorithms select clusters by an agglomerative rather than a divisive procedure, and the number of clusters to be examined must be stipulated. In the NORMIX algorithm convergence may be reached before the stipulated number of cluster configurations have been processed. For each program the recommended maximum number of clusters sought is seven.

Table II. Comparison of capabilities of the SAS CLUSTER, MIKCA and NORMIX algorithms.

Property	SAS	MIKCA	NORMIX
Complexity	Low	Medium	High
Output quantity	Minimum	Moderate	Extensive
Number of input data	250	500	2000[a]
Limit to number of variables	10	20	20[a]
Maximum number of clusters	250	15	150
Distance options	1	3	1
Criteria options	1	9	1
Hierarchical clustering	Yes with maps	No	Yes with maps

[a] May be increased if computer memory space permits. Computation time increases exponentially with the numbers of variates and observations.

Conclusions

The applications of three clustering algorithms to aerometric data bases were compared: SAS CLUSTER, MIKCA, and NORMIX. The three routines produce similar results through the first processing steps of hierarchical clustering and output of cluster means. Beyond that point MIKCA appears to provide slightly more information than SAS CLUSTER. NORMIX, as modified and expanded, produces considerably more information and guidance than either MIKCA or SAS CLUSTER.

Many other clustering programs may be used, but only the aforementioned three have been examined in this study, and only NORMIX has been examined in detail. Of the three, only NORMIX provides complete statistical estimates of the multimodal, multivariate normal distributions and provides probabilities of cluster membership for each datum. SAS CLUSTER is strictly hierarchical in grouping and mapping and uses this information as initial statistical estimates for further iterations to achieve maximum likelihood estimates. MIKCA is strictly hierarchical on the first pass of the data, but on successive iterations permits flow of data among the clusters to arrive at another cluster configuration which presumably is better.

Los Angeles Catalyst Study data were analyzed by use of the two algorithms, SAS CLUSTER and MIKCA. The results were similar. Community Health Air Monitoring Program data were analyzed by means of the expanded NORMIX program.

References

1. D. Randerson, ed., *Atmospheric Science and Power Production*, Technical Information Center, Office of Scientific and Technical Information, U.S. Department of Energy, DOE (TIC-27601)DE84005177, Washington, DC, 1984.
2. W. F. Hunt, Jr., R. B. Faoro, S. K. Goranson, "Ratio test and Shewhart control chart tests applied to the National Aerometric Data Bank," ASQC Tek. Con. 50–55 (1976).
3. K. Pearson, "Contributions to the mathematical theory of evolution," *Phil. Trans. Roy. Soc. London* 185: 71 (1894).
4. H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Ed. Psychol.* 24: 417, 489 (1933).
5. H. Hotelling, "Relations between two sets of variates," *Bimetrika* 28: 321 (1936).
6. H. Hotelling, "The relations of the newer multivariate statistical methods

to factor analysis," *B. J. Stat. Psychol.* **10:** 69 (1957).
7. J. E. Jackson, "Principal components," Part I, "Principal components," *J. Qual. Tek.* **12:** 201 (1980); Part II, "Additional topics related to principal components," *J. Qual. Tek.* **13:** 46 (1981); Part III, "What is factor analysis?," *J. Qual. Tek.* **13:** 125 (1981).
8. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, Inc., New York, 1958.
9. W. M. Cooley, P. R. Lohnes, *Multivariate Procedures for the Behavioural Sciences*, John Wiley and Sons, Inc., New York, 1962.
10. D. N. Lawley, A. E. Maxwell, *Factor Analysis as a Statistical Method*, Buterworth Mathematical Texts, London, 1963.
11. H. H. Harman, *Modern Factor Analysis*, 3rd Ed. Revised, University of Chicago Press, Chicago, IL, 1976.
12. D. F. Gatz, "Identification of aerosol sources in the St. Louis area," *J. Appl. Meteorol.* **17:** 600 (1978).
13. B. A. Roscoe, P. K. Hopke, S. L. Dattner, J. M. Jenks, "The use of principal component factor analysis to interpret particulate compositional data sets," *JAPCA* **32:** 637 (1982).
14. R. R. Sokal, P. H. A. Sneath, *Principles of Numerical Taxonomy*, W. H. Freeman and Co., San Francisco, 1963.
15. P. E. Green, J. D. Carrol, *Mathematical Tools for Applied Mathematical Analysis*, Academic Press, Inc., New York, 1976.
16. R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.* **7:** 179 (1936).
17. C. R. Rao, *Advanced Statistical Methods in Biometric Research*, John Wiley and Sons, Inc., New York, 1952.
18. R. C. Tryon, D. E. Bailey, *Cluster Analysis*, McGraw-Hill, Inc., New York, 1970.
19. M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, Inc., New York, 1973.
20. F. H. Marriott, *Cluster Analysis and Related Problems, The Interpretation of Multiple Observations*, Academic Press, Inc., New York, 1974.
21. J. A. Hartigan, *Clustering Algorithms*, John Wiley and Sons, Inc., 1975.
22. J. S. Fordyce, R. B. King, H. E. Neustadter, "Elemental composition of airborne particulates and source identification, data analysis techniques," *JAPCA* **26:** 1079 (1976).
23. R. O. Duda, P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, Inc., 1973.
24. B. Everitt, *Cluster Analysis*, Heinemann Educ. Books, London, 1977.
25. H. Spath, *Cluster Analysis Algorithms*, John Wiley and Sons, Inc., New York, 1980.
26. M. S. Aldenderfer, R. K. Blashfield, *Cluster Analysis*, Sage Publications, Beverly Hills, CA, 1984.
27. D. J. Hand, *Discrimination and Classification*, John Wiley and Sons, Inc., New York, 1981.
28. R. R. Hocking, "Developments in linear regression methodology: 1959–1982," *Technometrics* **25:** 219 (1983).
29. D. Pfeffermann, G. Nathan, "Regresion analysis of data from a cluster sample," *JASA* 681 (1976).
30. M. Goldstein, I. F. Goldstein, J. L. Fleiss, L. Landowitz, "Methodological problems arising from the choice of an independent variable in linear regression with application to an air pollution epidemiological study," *Environ. Health Perspect.* 311 (1979).
31. M. Abramowitz, I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, U.S. Department of Commerce, NBS, Applied Mathematics Series 55, 1968.
32. J. H. Wolfe, "Computation Methods for Estimating the Parameters of Multivariate Normal Mixture of Distributions," Research Memorandum SRM 68-2, *U.S. Naval Pers. Research Facility*, San Diego, CA, 1967.
33. J. H. Wolfe, "NORMIX 360 Computer Program," Research Memorandum SRM 72-4, *U.S. Naval Pers. and Training Res. Lab.*, San Diego, CA, 1971.
34. D. J. McRae, "MIKCA: A FORTRAN iterative K-means cluster analysis program," *Behavioral Science* **16:** 123 (1971); rev. M. J. Symons (1973).
35. A. J. Barr, J. H. Goodnight, J. P. Sale, J. T. Helweg, *A User's Guide to SAS*, Spanks Press, Raleigh, NC, 1976.
36. R. C. Rhodes, "Application of Cluster Analysis to Aerometric Data," Vol. 1, Part 1, "Clustering, validation and classification of data," Part 2, "Investigation and report of cluster analysis," (1980); Vol. 2, "Modifications and options applied to Wolfe's NORMIX 360 cluster analysis program," M. E. Graves, (1980); Vol. 3, "Separation of environmental data into clusters by the NORMIX program," 80 (1980); Project Summary, "Applications of cluster analysis to aerometric data," 8 (1980); National Technical Information Service (NTIS), Springfield VA.
37. H. L. Crutcher, "On the standard vector deviation wind rose," *J. Meteorol.* **14:** 28 (1957).
38. J. Aitchison, J. A. C. Brown, *The Lognormal Distribution*, Cambridge University Press, New York, 1976.
39. P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. Nat. Inst. Sci. India* **2:** 1 (1936).
40. J. H. Ward, Jr., "Hierarchical grouping to optimise an objective function," *J. Am. Stat. Assoc.* **58:** 236 (1963).
41. J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics, and Probability*, AD669871, University of California Press, Berkeley, CA, 1967.
42. J. E. Gentle, *Current Index to Statistics, Applications and Theory*, American Statistical Association, Institute of Mathematics and Statistics, Washington, DC, 1984.
43. J. C. Bezdek, C. Coray, R. Gunderson, J. Watson, "Detection and characterization of cluster sub-structure, I, linear structure, fuzzy c-lines," *SIAM APMA* **40:** 339 (1981); "Detection and characterization of cluster substructure, II, fuzzy c-variates and convex combinations thereof," *SIAM APMA* **40:** 358 (1981).
44. P. Smets, "The degree of belief in a fuzzy event," *Infor. Sci.* **25:** 1 (1981).
45. E. Anderson, "The irises of the Gaspe Peninsula," *Bull. Am. Iris Soc.* **59:** 2–5 (1935).
46. F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics* **11:** 1 (1969).
47. V. Barnett, T. Lewis, *Outliers in Statistical Data*, John Wiley and Sons, New York, 1978.

Dr. Harold Crutcher is a certified consulting meteorologist (AMS-81) and certified reliability engineer (ASQC-140), 35 Westall Avenue, Asheville, NC 28804. R. C. Rhodes is a quality assurance specialist, U.S. EPA, MD-77B Research Triangle Park, NC 27711. Maurice E. Graves is a project scientist, Northrop Services, Inc., P.O. Box 12313, Research Triangle Park, NC 27709. A. Carl Nelson is currently a Professor in the Department of Mathematical Sciences, University of North Carolina at Wilmington, P.O. Box 3725, Wilmington, NC 28406. He was previously branch manager, PEDCo Environmental, Inc. (now PEI Associates, Inc.), 505 S. Duke Street, Suite 503, Durham, NC 27701. Beth Fairbairn was previously a data analyst, PEDCo Environmental, Inc., Durham, NC 27701. This paper was submitted for peer review May 24, 1982; the revised manuscript was received April 24, 1985.